ORIGINAL ARTICLE

DOI: 10.4274/hamidiyemedj.galenos.2022.15238

Hamidiye Med J 2022;3(3):162-170

Investigation of the Change in the Correct Classification Ratios by Using the Richard Link Function in Logistic Regression: A Research on the Determination of Risk Factors in COPD

Lojistik Regresyonda Richard Link Fonksiyonu Kullanımı ile Doğru Sınıflama Oranlarındaki Değişimin İncelenmesi: KOAH'da Risk Faktörlerinin Belirlenmesi Üzerine Bir Araştırma

Kürşad Nuri Baydili[1], Mustafa Çörtük[2], Ahmet Dirican[1]

[1]*İstanbul University-Cerrahpaşa, Cerrahpaşa Faculty of Medicine, Department of Biostatistics, İstanbul, Türkiye*

[2]*University of Health Sciences Türkiye Hamidiye Faculty of Medicine; İstanbul Yedikule Chest Diseases and Thoracic Surgery Training and Research Hospital, Clinic of Chest Diseases and Tuberculosis, İstanbul, Türkiye*

ABSTRACT

Background: Regression analyses are used to explain the relationship between a dependent variable and independent variables using mathematical models. Logistic regression, which is used in cases where the dependent variable is categorical, is often used in analyzing health-related data.

Materials and Methods: It is well known that the inflection point of the logistic regression curve sometimes corresponds to smaller or larger values on the horizontal axis, resulting in incorrect classifications. The present study aimed to increase correct classification rates by using the Richards link function to determine the most suitable inflection point for data.

Results: In order to evaluate the performance of the Richards link function, four different simulation scenarios and applications were carried out with a total of 1.005 individuals, of whom 505 were non-chronic obstructive pulmonary disease (COPD) and 500 were COPD individuals. The data were divided into learning and test data. A logistic regression model was obtained from the learning data, and an increase in the correct classification rates was observed with the use of the Richards link function in this model. The model was applied to the test data with the m-value determined for the learning data set and achieved a higher correct classification rate than the current method.

Conclusion: The present study indicated that certain percentage increases can be achieved in correct classification rates by using the Richards link function. However, it would be beneficial to conduct studies in which applications are made with data sets containing fewer and more independent variables, different sample sizes, and combinations of independent variable types.

Keywords: Logistic regression, Richard link function, correct classification

ÖZ

Amaç: Bir bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin matematiksel modeller kullanılarak açıklanması için regresyon analizleri kullanılır. Bağımlı değişkenin kategorik olduğu durumlarda kullanılan lojistik regresyon özellikle sağlık alanında sıklıkla kullanılır.

Gereç ve Yöntemler: Lojistik regresyon eğrisinin bükülme noktasının yatay eksende bazen olması gerekenden daha küçük ya da daha büyük değerlere karşılık geldiği, bunun sonucunda da hatalı sınıflandırmalar yaptığı bilinmektedir. Araştırmada; verilere en uygun büküm noktasının tespiti için Richard link fonksiyonu kullanılarak doğru sınıflama oranlarında artışlar sağlanması hedeflenmiştir.

Bulgular: Richard link fonksiyonunun performansını değerlendirmek amacıyla; dört farklı simülasyon senaryosu ve 505'i kronik obstrüktif akciğer hastalığı (KOAH) olmayan 500'ü ise KOAH olan toplamda 1,005 bireyden oluşan gerçek verilerle uygulamalar gerçekleştirilmiştir. Gerçek verilerle gerçekleştirilen uygulamalarda ise tüm verilerin kullanılmasıyla elde edilen modelde Richard link fonksiyonu kullanımı ile birlikte doğru sınıflama oranlarında artışlar olduğu saptanmıştır. Verilerin öğrenme ve test verileri

Address for Correspondence: Kürşad Nuri Baydili, İstanbul University-Cerrahpaşa, Cerrahpaşa Faculty of Medicine, Department of Biostatistics, İstanbul, Türkiye
Phone: +90 541 739 23 23 E-mail: kursadnuri.baydili@sbu.edu.tr **ORCID ID:** orcid.org/0000-0002-2785-0406
Received: 06.09.2022 **Accepted:** 26.09.2022

ÖZ

olarak ikiye ayrılmış, öğrenme verilerinden lojistik regresyon modeli elde edilmiş, bu modelde de Richard link fonksiyonu kullanımı ile birlikte doğru sınıflama oranlarında artış gözlemlenmiştir. Öğrenme veri seti için belirlenen m değeri için model test verilerine uygulanmış ve mevcut yöntemden daha yüksek doğru sınıflama oranına ulaştığı görülmüştür.

Sonuç: Richard link fonksiyonu kullanılarak doğru sınıflama oranlarında belirli oranlarda artışlar sağlanabileceği görülmüştür. Ancak, daha az ve daha çok sayıda bağımsız değişkenler içeren, farklı örneklem büyüklüklerinde ve bağımsız değişken türlerinde kombinasyonların denendiği veri setleriyle uygulamaların yapıldığı çalışmaların yapılmasında fayda görülmektedir.

Anahtar Kelimeler: Logistik regresyon, Richard link fonksiyonu, doğru sınıflama

Introduction

Establishing a cause-effect relationship between the dependent variable and the independent variable(s) is one of the aims of scientific research (1). Univariate methods used when investigating causality make the comparison by assuming that all other factors other than the variable are homogeneous or constant. However, sometimes it is not possible to achieve homogeneity or stability in the real world. A variable can often change with one or more variables. The problem can be solved by including these co-variations through multivariate statistical analysis (2), which aims to predict an outcome with multiple independent variables (3).

Multivariate statistical analysis includes many techniques depending on the purpose of the study, the type of dependent and independent variable(s), and the fulfillment of certain conditions (2). Regression analysis is used to estimate relationships between a dependent variable and a set of independent variables using mathematical models (4). Researchers in the field of health generally aim to classify their observations and make inferences about future observations based on existing observations (5). The first known classification methods were cluster analysis, which was originated by Driver and Kroeber (6) in social and life sciences, and two-group discriminant analysis proposed by Fisher (7). The primary techniques used to classify observations are cluster analysis, discriminant analysis, and logistic regression analysis (8). In cluster analysis, where the number of groups is unknown, data are assigned to groups according to certain criteria (9). In discriminant analysis and logistic regression, although the number of groups is known, data are assigned to groups by using this information (8). In logistic regression, assumptions such as normality and homogeneity of variances required in discriminant analysis are not sought (4).

Berkson (10) was the first to publish an application of the logistic model in the field of biology. Logistic regression aims to reveal the model that has the highest fit with the least number of variables (9). Logistic regression can be used to estimate and summarize data, as well as for classification by examining the relationship between the dependent variable and the independent variable(s). Logistic regression is used when the dependent variable is in the form of qualitative data (4). There are three different types of logistic regression: Binary logistic regression is used when the dependent variable has two categories, multinomial logistic regression is used when the dependent variable has more than two categories, and ordinal logistic regression is used when the dependent variable is measured at the ordinal level (2,4,11).

Logistic regression is commonly used in such fields as economics, education and health (12). Binary logistic regression has become an increasingly employed statistical tool in medical research, and is generally concerned with whether there is a risk, such as disease (13), and is coded as 1 and 0. An odds ratio is used in risk estimation in retrospective studies. The significance of the odds ratio is determined by examining the confidence intervals. If the confidence interval for the odds ratio does not include the number 1, then the calculated odds ratio is considered statistically significant. If the calculated odds ratio is found to be significant, an odds ratio greater than 1 indicates that the factor is a risk factor, while an odds ratio of less than 1 indicates that it is a protective factor (4). In binary logistic regression, a logistic regression model is created [$\pi(x)$] by calculating the probability of Y being 1 $P(Y=1|X=x)$ when the value of the independent variable (X) is known.

The logistic regression graph is an S-shaped sigmoid curve (14,15). The logistic curve was first used by Verhulst (16) to describe the growth in a population. The inflection point of this curve may sometimes correspond to smaller or larger x-values than it should be. Gürcan et al. (17) stated that in such cases, by using various link functions, the x-values corresponding to the inflection point of the curve may be more optimal, and thus, there may be an increase in the rates of correct classification. They found the inflection point by analyzing its second derivative of the curve proposed by Richards (18), and aimed to increase the correct classification rate of the model by applying the inflection points separately for the misclassified observations.

Material and Methods

Ethics committee approval with the number 18/1 was obtained from the Hamidiye Scientific Research Ethics Committee at the meeting numbered 2022/18 for the research. The present study aimed to increase the correct classification rate by using an alternative link function to the existing method used in binary logistic regression. In logistic regression, instead of the $P = \frac{e^a}{1+e^a}$ forwmula, the model was changed with 0.01 increments in the m (1,3) interval, the probability values were calculated using the Richards link function with the $P = (1 + (m-1).(e^{-a}))^{\frac{1}{1-m}}$ formula, and the estimated classification values were obtained according to these probability values. Data were collected through face-to-face interviews using a questionnaire for a total of 1005 individuals, 505 without chronic obstructive pulmonary (COPD) and 500 with COPD, to evaluate the performance of the Richards link function. The demographic characteristics of the study participants are presented in Table 1. Applications were carried out in two different ways. In the first method, all the data were included in the logistic regression model and the probability and class values were obtained with the proposed method. Then, probability calculations were made using the Richards link function with the same coefficients. Subsequently, the m-value, which maximizes the correct classification percentage, was determined. In the second method, 74.6% (n=750) of the data were included in the logistic regression model, and the correct classification numbers and ratios for all m-values were presented with the coefficients. Next, probability and classification values were obtained for the remaining 25.4% (n=255) of the data using the m-value, which maximizes the correct classification percentage.

Table 1. Demographic information of participants

	Learning data set	Test data set	Total
	n (%)	n (%)	n (%)
Gender			
Male	461 (61.5)	176 (69)	637 (63.4)
Female	289 (38.5)	79 (31)	368 (36.6)
Packs of cigarettes smoked per year>10 per year			
No	406 (54.1)	149 (58.4)	555 (55.2)
Yes	344 (45.9)	106 (41.6)	450 (44.8)
COPD in relatives			
No	495 (66)	184 (72.2)	679 (67.6)
Yes	255 (34)	71 (27.8)	326 (32.4)
Lung disease other than COPD in relatives			
No	575 (76.7)	190 (74.5)	765 (76.1)
Yes	175 (23.3)	65 (25.5)	240 (23.9)
Place of residence			
Other	227 (30.3)	104 (40.8)	331 (32.9)
Metropolis	523 (69.7)	151 (59.2)	674 (67.1)
Duration of daily exercise>1 hour			
≤60 min	607 (80.9)	244 (95.7)	851 (84.7)
>60 min	143 (19.1)	11 (4.3)	154 (15.3)
COPD			
No	375 (50)	130 (51)	505 (50.2)
Yes	375 (50)	125 (49)	500 (49.8)
	Med (min-max)	**Med (min-max)**	**Med (min-max)**
Age	52.5 (14-91)	52 (16-94)	52 (14-94)
BMI	26.40 (15.59-45.71)	25.54 (15.62-57.11)	26.13 (15.59-57.11)
Duration of exposure to wood, dung or coal smoke	0 (0-50)	0 (0-70)	0 (0-70)

COPD: Chronic obstructive pulmonary disease, BMI: Body mass index

Results

In the data set in which all observations were included, the variables were individually included in the logistic regression model to select the variables suitable for the logistic regression model. It was concluded that the variables of gender (p<0.001), age (p<0.001), body mass index (p<0.001), duration of exposure to wood, dung or coal smoke (p<0.001), smoking status over 10 packs/year (p<0.001), having a relative with COPD (p<0.001), having a recent lung disease other than COPD (p<0.001), place of residence (p<0.001), and daily exercise for more than 1 hour (p<0.001) should be included in the model (Table 2).

Logistic regression in which all variables were included in the model showed that the variables of gender (p=0.946) and body mass index (p=0.307) had no effect on COPD status. It was found that a 1-unit increase in age was a 1.148-fold greater risk (p<0.001), and a 1-unit increase in exposure time to wood, dung or coal smoke was a 1.027-fold greater risk (p=0.011). It was determined that smoking more than 10 packs/year was a 7.832-fold greater risk (p<0.001), having COPD in first-degree relatives was a 2.792-fold greater risk (p<0.001), having individuals with lung disease other than COPD in first-degree relatives was a 4.068-fold greater risk (p<0.001), living in a metropolis was a 7.664-fold greater risk (p<0.001), and exercising for more than 1 hour a day was a 0.04-fold greater risk (25-fold protective factor) (p<0.001) (Table 3).

The correct classification rate obtained with the available variables was 93% (n=935) (Table 4). By using the Richards link function in probability calculations, it was found that the correct classification rate for the 10 value of m in the range (1.42,1.51) was higher than the correct classification rates for the other m-values. The results of the observations with changes in the estimated classification values for m=1.42 are presented in Table 5.

By using the Richards link function, for 6 observations with a change in classification values for m=1.42, the m-values that gave the highest correct classification rate for values varying in the range (1,6) were determined. It was determined that the current method gave the correct result

Table 2. Univariate logistic regression results

	B	S.E.	Wald	p	OR (95% CI)
Gender (ref: Male)	-1.117	0.137	66.06	<0.001*	0.327 (0.25-0.428)
Age	0.167	0.011	241.765	<0.001*	1.182 (1.157-1.207)
BMI	0.105	0.016	42.175	<0.001*	1.111 (1.076-1.147)
Duration of exposure to wood, dung or coal smoke	0.088	0.007	153.198	<0.001*	1.092 (1.077-1.107)
Packs of cigarettes smoked per year>10 per year (ref: No)	2.716	0.159	292.175	<0.001*	15.12 (11.074-20.645)
COPD in relatives (ref: None)	1.289	0.145	79.451	<0.001*	3.629 (2.734-4.818)
Lung disease other than COPD in relatives (ref: None)	0.94	0.155	36.594	<0.001*	2.561 (1.888-3.473)
Place of residence (ref: Metropolis)	2.034	0.161	159.907	<0.001*	7.643 (5.576-10.475)
Duration of daily exercise>1 hour (ref: No)	-3.364	0.393	73.276	<0.001*	0.035 (0.016-0.075)

*p<0.05, OR: Odds ratio, CI: Confidence interval, COPD: Chronic obstructive pulmonary disease, S.E.: Standard error

Table 3. Multivariate logistic regression table

	B	S.E.	Wald	p	OR (95% CI)
Gender (ref: Male)	-0.02	0.299	0.005	0.946	0.98 (0.545-1.76)
Age	0.138	0.013	110.701	<0.001*	1.148 (1.119-1.178)
BMI	-0.034	0.033	1.046	0.307	0.967 (0.907-1.031)
Duration of exposure to wood, dung or coal smoke	0.027	0.011	6.402	0.011*	1.027 (1.006-1.048)
Packs of cigarettes smoked per year>10 per year (ref: No)	2.058	0.303	46.135	<0.001*	7.832 (4.325-14.184)
COPD in relatives (ref: None)	1.027	0.281	13.349	<0.001*	2.792 (1.61-4.844)
Lung disease other than COPD in relatives (ref: None)	1.403	0.322	18.989	<0.001*	4.068 (2.164-7.646)
Place of residence (ref: Metropolis)	2.037	0.319	40.702	<0.001*	7.664 (4.1-14.329)
Duration of daily exercise>1 hour (ref: No)	-3.207	0.551	33.834	<0.001*	0.04 (0.014-0.119)
Constant	-7.65	1.094	48.916	<0.001*	0

*p<0.05, S.E.: Standard error, COPD: Chronic obstructive pulmonary disease, BMI: Body mass index, OR: Odds ratio, CI: Confidence interval

in two of these 6 observations, while the proposed method gave the correct result in four of them. With the proposed method, it was observed that an increase of approximately 0.2% (n=2) occurred in the correct classification rates (Table 5).

In the selection of the variables suitable for the logistic regression model, the variables were examined one by one by including them in the logistic regression model. It was concluded that the following variables should be included in the model: Gender (p<0.001), age (p<0.001), body mass index (p<0.001), duration of exposure to wood, dung or coal smoke (p<0.001), smoking status over 10 packs/year (p<0.001), having a relative with COPD (p<0.001), having a relative with lung disease other than COPD (p<0.001), place of residence (p<0.001), and exercising for more than 1 hour daily (p<0.001) (Table 6).

The logistic regression model showed that gender (p=0.727) and body mass index (p=0.643) had no effect on having COPD. It was determined that a 1-unit increase in age was a 1.195-fold greater risk (p<0.001) and 1-unit increase in exposure time to wood, dung or coal smoke was 1.069-fold greater risk for having COPD (p<0.001). It was determined that smoking more than 10 packs/year was a 16.446-fold greater risk (p<0.001), having COPD in first-degree relatives was a 3.348-fold greater risk (p=0.002), having individuals with lung disease other than COPD in first-degree relatives was a 9.797-fold greater risk, living in a metropolitan area was a 17.288-fold greater risk (p<0.001), and exercising for more than 1 hour a day was a 71.43-fold protective factor (p<0.001) (Table 7).

Table 4. Classification table of logistic regression equation

	Estimated		Correct classification rate
Observed	**Absence of COPD**	**Presence of COPD**	
Absence of COPD	466	39	92.3
Presence of COPD	31	469	93.8
			93.0

COPD: Chronic obstructive pulmonary disease

Table 5. Observations with changes in estimated classification values for m=1.42

p	Class	$p_{m=1.42}$	$Class_{m=1.42}$	Observed
0.521	1	0.460	0	1
0.540	1	0.483	0	0
0.540	1	0.483	0	0
0.507	1	0.442	0	1
0.545	1	0.489	0	0
0.538	1	0.480	0	0

Table 6. Univariate logistic regression results in learning data set

	B	S.E.	Wald	p	OR (95%CI)
Gender (ref: Male)	-1.404	0.161	75.89	<0.001*	0.246 (0.179-0.337)
Age	0.195	0.015	165.484	<0.001*	1.215 (1.179-1.251)
BMI	0.121	0.019	38.738	<0.001*	1.128 (1.086-1.172)
Duration of exposure to wood, dung or coal smoke	0.124	0.01	155.665	<0.001*	1.132 (1.11-1.154)
Packs of cigarettes smoked per year >10 per year (ref: No)	2.866	0.187	233.796	<0.001*	17.563 (12.164-25.36)
COPD in relatives (ref: None)	1.531	0.17	81.25	<0.001*	4.622 (3.313-6.447)
Lung disease other than COPD in relatives (ref: None)	1.041	0.184	32.023	<0.001*	2.832 (1.975-4.061)
Place of residence (ref: Metropolis)	2.007	0.192	109.004	<0.001*	7.438 (5.103-10.84)
Duration of daily exercise >1 hour (ref: No)	-3.567	0.425	70.333	<0.001*	0.028 (0.012-0.065)

*p<0.05, COPD: Chronic obstructive pulmonary disease, BMI: Body mass index, S.E.: Standard error, OR: Odds ratio, CI: Confidence interval

The correct classification rate of the equation with the available variables was 94.5% (n=709) (Table 8). By using the Richards link function in probability calculations, it was determined that the correct classification rate for 23 different values of m in the (1.33; 1.44) and (1.56; 1.66) ranges was higher than the correct classification rates for the other m-values. The results of the observations with changes in the estimated classification values for m=1.66 are presented in Table 9.

By using the Richards link function, 6 observations with changes in classification values for m=1.66 were determined from 23 different m-values, which gave the highest correct classification rate for values varying in the (1,3) range. The current method gave the correct result in one of these six observations, while the proposed method gave the correct results in five of these six observations. It was observed that there was an increase of approximately 0.67% (n=5) in the correct classification rate with the proposed method (Table 9).

The application of the models from the training data to the test data gave the following results: Correct classification was made for 221 observations with the current method and 222 observations with the proposed method. The probability value calculated with the current method was found to be above 0.5, while the probability values calculated by the proposed method for the observation estimated to be COPD were found to be below 0.5. The correct classification was made in the direction of not having COPD. As a result, the model, which increased the rate of correct classification by 0.67% in the training data, also provided an increase in correct classification of approximately 0.4% in the test data (Table 10).

Table 7. Multivariate logistic regression table in learning data set

	B	S.E.	Wald	p	OR (95% CI)
Gender (ref: Male)	0.145	0.415	0.122	0.727	1.156 (0.513-2.607)
Age	0.178	0.023	61.601	**<0.001***	1.195 (1.143-1.249)
BMI	-0.022	0.047	0.215	0.643	0.978 (0.892-1.073)
Duration of exposure to wood, dung or coal smoke	0.067	0.019	12.841	**<0.001***	1.069 (1.031-1.109)
Packs of cigarettes smoked per year >10 per year (ref: No)	2.8	0.437	41.115	**<0.001***	16.446 (6.988-38.705)
COPD in relatives (ref: None)	1.208	0.394	9.39	**0.002***	3.348 (1.546-7.251)
Lung disease other than COPD in relatives (ref: None)	2.282	0.51	19.992	**<0.001***	9.797 (3.603-26.639)
Place of residence (ref: Metropolis)	2.85	0.479	35.416	**<0.001***	17.288 (6.762-44.195)
Duration of daily exercise >1 hour (ref: No)	-4.281	0.712	36.153	**<0.001***	0.014 (0.003-0.056)
Constant	-10.701	1.737	37.932	**<0.001***	0

*p<0.05, OR: Odds ratio, CI: Confidence interval, COPD: Chronic obstructive pulmonary disease, BMI: Body mass index, S.E.: Standard error

Table 8. Classification table of logistic regression equation in learning data set

	Estimated		Correct classification rate
Observed	**Absence of COPD**	**Presence of COPD**	
Absence of COPD	353	22	94.1
Presence of COPD	19	356	94.9
			94.5

COPD: Chronic obstructive pulmonary disease

Table 9. Observations with changes in the estimated classification values for m=1.66

p	Class	$p_{m=1.42}$	$Class_{m=1.66}$	Observed
0.506	1	0.471	0	1
0.532	1	0.499	0	0
0.516	1	0.481	0	0
0.502	1	0.466	0	0
0.519	1	0.485	0	0
0.504	1	0.468	0	0

Table 10. Probability and classification values obtained as a result of applying the model to the test data

p	Yp	C	p	Yp	C	p	Yp	C	p	Yp	C	p	Yp	C
0.996	0.996	1	0.994	0.994	1	0.984	0.984	1	0.290	0.233	0	<0.001	<0.001	0
1.000	1.000	1	1.000	1.000	1	0.937	0.936	1	0.688	0.673	0	<0.001	<0.001	0
0.997	0.997	0	1.000	1.000	1	0.889	0.887	1	0.003	<0.001	0	<0.001	<0.001	0
1.000	1.000	1	0.949	0.948	1	0.997	0.997	1	0.845	0.841	0	<0.001	<0.001	0
1.000	1.000	1	0.979	0.979	1	0.992	0.991	1	0.010	0.002	0	<0.001	<0.001	0
0.999	0.999	1	0.994	0.994	1	0.258	0.199	0	<0.001	<0.001	0	<0.001	<0.001	0
1.000	1.000	1	0.995	0.995	1	1.000	1.000	1	<0.001	<0.001	0	<0.001	<0.001	0
1.000	1.000	1	0.690	0.675	1	0.965	0.965	1	0.047	0.018	0	0.001	<0.001	0
0.997	0.997	1	1.000	1.000	1	0.707	0.693	0	<0.001	<0.001	0	<0.001	<0.001	0
0.994	0.994	1	0.978	0.978	1	0.965	0.965	1	0.001	<0.001	0	<0.001	<0.001	0
0.985	0.985	1	1.000	1.000	1	0.636	0.616	1	0.003	<0.001	0	0.002	<0.001	0
0.902	0.900	1	1.000	1.000	1	0.984	0.984	0	<0.001	<0.001	0	0.001	<0.001	0
0.997	0.997	1	0.636	0.616	1	0.794	0.787	1	<0.001	<0.001	0	0.161	0.104	0
0.032	0.010	1	0.695	0.681	1	0.999	0.999	1	0.169	0.112	0	<0.001	<0.001	0
0.393	0.345	1	0.997	0.997	1	0.999	0.999	1	0.004	<0.001	0	0.341	0.288	0
0.970	0.970	1	0.944	0.944	1	0.999	0.999	1	0.001	<0.001	0	<0.001	<0.001	0
1.000	1.000	1	0.971	0.971	1	0.982	0.981	1	0.052	0.020	0	0.002	<0.001	0
0.999	0.999	1	0.988	0.988	1	0.999	0.999	1	<0.001	<0.001	0	<0.001	<0.001	0
0.973	0.973	0	0.992	0.992	1	0.998	0.998	1	<0.001	<0.001	0	<0.001	<0.001	0
0.990	0.990	0	0.949	0.949	1	0.261	0.202	1	<0.001	<0.001	0	0.001	<0.001	0
0.853	0.849	0	1.000	1.000	1	0.001	<0.001	1	<0.001	<0.001	0	<0.001	<0.001	0
0.456	0.414	1	0.998	0.998	1	0.984	0.984	1	0.073	0.033	0	0.002	<0.001	0
0.974	0.974	1	0.935	0.934	1	0.998	0.998	1	<0.001	<0.001	0	<0.001	<0.001	0
0.963	0.963	1	0.695	0.681	1	0.999	0.999	1	<0.001	<0.001	0	0.003	<0.001	0
0.996	0.996	0	1.000	1.000	1	0.998	0.998	1	0.002	<0.001	0	0.600	0.575	0
0.024	0.006	0	0.999	0.999	1	0.998	0.998	1	<0.001	<0.001	0	<0.001	<0.001	0
0.757	0.748	1	1.000	1.000	1	0.890	0.888	1	<0.001	<0.001	0	0.007	0.001	0
0.003	0.000	1	0.002	<0.001	0	0.766	0.757	1	0.001	<0.001	0	0.002	<0.001	0
0.828	0.823	1	0.999	0.999	1	0.999	0.999	1	<0.001	<0.001	0	0.021	0.005	0
0.970	0.970	1	0.984	0.984	1	0.997	0.997	1	<0.001	<0.001	0	0.029	0.009	0
0.011	0.002	1	0.994	0.994	1	0.997	0.997	1	0.001	<0.001	0	0.045	0.017	0
0.994	0.994	1	0.270	0.212	0	0.997	0.997	1	0.001	<0.001	0	0.009	0.002	0
0.950	0.949	1	0.849	0.845	1	0.917	0.916	1	<0.001	<0.001	0	<0.001	<0.001	0
0.999	0.999	1	0.993	0.993	1	0.999	0.999	1	<0.001	<0.001	0	<0.001	<0.001	0
0.005	0.001	0	0.997	0.997	1	0.980	0.980	1	<0.001	<0.001	0	0.883	0.881	0
0.846	0.842	1	0.006	0.001	1	0.996	0.996	1	0.054	0.021	0	0.955	0.955	0
0.996	0.996	1	0.993	0.993	1	<0.001	<0.001	0	<0.001	<0.001	0	0.922	0.921	0
0.999	0.999	1	0.999	0.999	1	<0.001	<0.001	0	<0.001	<0.001	0	0.836	0.832	0
0.889	0.887	1	0.969	0.969	1	0.062	0.027	0	<0.001	<0.001	0	0.934	0.933	0
0.948	0.948	1	0.080	0.038	1	0.005	0.001	0	<0.001	<0.001	0	0.638	0.618	0
0.936	0.935	1	0.999	0.999	1	<0.001	<0.001	0	0.001	<0.001	0	0.094	0.048	0
0.985	0.985	1	0.596	0.571	1	0.035	0.011	0	<0.001	<0.001	0	0.314	0.259	0

Table 10. Continue

p	Yp	C	p	Yp	C	p	Yp	C	p	Yp	C	p	Yp	C
0.982	0.982	**1**	1.000	1.000	**1**	0.232	0.172	**0**	<0.001	<0.001	**0**	0.010	0.002	**0**
0.828	0.823	**1**	0.970	0.970	**1**	0.017	0.004	**0**	<0.001	<0.001	**0**	0.003	<0.001	**0**
0.992	0.992	**1**	0.005	0.001	**0**	0.989	0.989	**0**	<0.001	<0.001	**0**	0.008	0.001	**0**
0.996	0.996	**1**	0.896	0.894	**1**	0.672	0.655	**0**	<0.001	<0.001	**0**	0.895	0.893	**0**
0.318	0.263	**1**	0.969	0.969	**1**	0.001	<0.001	**0**	0.025	0.007	**0**	0.104	0.056	**0**
0.498	0.462	**1**	0.983	0.983	**1**	<0.001	<0.001	**0**	0.011	0.002	**0**	0.096	0.050	**0**
0.989	0.989	**1**	1.000	1.000	**1**	<0.001	<0.001	**0**	0.001	<0.001	**0**	0.942	0.942	**0**
0.961	0.961	**1**	0.999	0.999	**1**	<0.001	<0.001	**0**	<0.001	<0.001	**0**	0.526	0.493	**0**
0.001	0.000	**1**	0.892	0.890	**1**	0.145	0.090	**0**	<0.001	<0.001	**0**	0.941	0.941	**0**

C: Status of having chronic obstructive pulmonary disease
p: Probability value obtained with current application
Yp: Probability value obtained with proposed application

Discussion

Human beings have been trying for centuries to instill some human skills in inanimate beings (19). In the 20th century, there has been an increasing interest in this subject among scientists. At the beginning of the second half of the 20th century, the Turkish scientist Arf (20) raised the question "Can Machines Think and How Can They Think?". One of the tasks undertaken by artificial intelligence is to give machines the ability to make inferences and decisions based on past experiences (21). As in many fields, in the field of health, researchers aim to make inferences about future observations with the data of existing observations by classifying these observations (5). Methods such as discriminant analysis, cluster analysis, and logistic regression are some of the methods used for classification (8). Logistic regression is mostly used when the classes of observations are known (22). Many studies have been carried out to improve the predictions made with logistic regression and increase the rate of correct classification (12,23,24). The logistic regression plot is an S-shaped sigmoid curve (25). The inflection point of this curve may sometimes take smaller or larger values than it should be in logistic regression. Gürcan et al. (17) stated that if the inflection point of the logistic curve corresponds to smaller or larger x-values than it should be, a more optimal inflection point can be determined by using various link functions. They found the inflection point by analyzing its second derivative of the curve proposed by Richards (18), and aimed to increase the correct classification rate of the model by applying the inflection points separately for the misclassified observations. This study aimed to make the inflection point of the logistic regression curve more ideal by using the link function $\pi(x) = (1 + (m-1).(e^{-a}))^{\frac{1}{1-m}}$ equation proposed by Richards (18). Probability values were calculated separately for all observations and assigned to classes according to these values. Then, the m-values that maximized the correct classification rate of the model were determined. Application of the model with all data indicated that the percentage of correct classification, which was 93% with the current method, increased by approximately 0.2% for m=1.42 using the Richards link function. Then, the data were split into training (n=750) and test (n=255) data sets. In the training data set, the correct classification rate, which was 94.5% with the current method, was found to be 95.07% for 23 different m-values in the (1.33; 1.44) and (1.56; 1.66) intervals using the Richards link function. For these values, the same model was applied to the test data for m=1.66. Correctly classifying 1 observation that was misclassified by the current method provided an increase of approximately 0.4% in the correct classification rate for the test data.

Conclusion

The present study was carried out to make predictions with a higher percentage of correct classification in logistic regression. It can be concluded that certain percentage increases in the correct classification rates can be achieved by using the Richards link function. However, it would be beneficial to carry out studies in which simulation scenarios are made with data sets containing fewer and more independent variables, different sample sizes, and combinations of independent variable types.

Information: This study is derived from the thesis study of Kürşad Nuri Baydili, PhD student of İstanbul University-Cerrahpaşa, Cerrahpaşa Faculty of Medicine, Department of Biostatistics.

Ethics

Ethics Committee Approval: Ethics committee approval with the number 18/1 was obtained from the Hamidiye Scientific Research Ethics Committee at the meeting numbered 2022/18 for the research.

Informed Consent: Retrospective study.

Peer-review: Internally and externally peer-reviewed.

Authorship Contributions

Surgical and Medical Practices: A.D., Concept: K.N.B., M.Ç., A.D., Design: K.N.B., M.Ç., A.D., Data Collection or Processing: K.N.B., M.Ç., A.D., Analysis or Interpretation: K.N.B., A.D., Literature Search: K.N.B., A.D., Writing: K.N.B., M.Ç.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Karagöz Y. SPSS AMOS META Uygulamalı Nitel-Nicel Karma Bilimsel Araştırma Yöntemleri ve Yayın Etiği, 2. Basım, Ankara: Nobel Yayıncılık; 2019. [Crossref]
2. Özdamar K. Paket Programlar ile İstatistiksel Veri Analizi 2, Kaan Kitabevi, Eskişehir; 2010. [Crossref]
3. Katz MH. Multivariable analysis: a practical guide for clinicians and public health researchers. Cambridge university press, 2011. [Crossref]
4. Alpar R. Uygulamalı çok değişkenli istatistiksel yöntemler. Detay Yayıncılık, 2017. [Crossref]
5. Karakoyun M, Hacıbeyoğlu M. Biyomedikal Veri Kümeleri İle Makine Öğrenmesi Sınıflandırma Algoritmalarinin İstatistiksel Olarak Karşılaştırılması. Mühendislik Bilimleri Dergisi. 2014;16:30-42. [Crossref]
6. Driver HE, Kroeber AL. Quantitative expression of cultural relationships (vol. 31, no. 4). Berkeley: University of California Press; 1932. [Crossref]
7. Fisher RA. The use of multiple measurements in taxonomic problems. Annals of eugenics, 1936;7:179-188. [Crossref]
8. Tatlıdil H. Uygulamalı Çok Değişkenli İstatistiksel Analiz. Cem Web Ofset, 1996. [Crossref]
9. Bircan H. Lojistik regresyon analizi: Tıp verileri üzerine bir uygulama. Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi. 2004;2:185-208. [Crossref]
10. Berkson J. Application of the logistic function to bio-assay. Journal of the American Statistical Association. 1994;39:357-365. [Crossref]
11. Şenel S, Alatlı B. Lojistik regresyon analizinin kullanıldığı makaleler üzerine bir inceleme. Journal of Measurement and Evaluation in Education and Psychology. 2014;5:35-52. [Crossref]
12. Sancar N, Inan D. A new alternative estimation method for Liu-type logistic estimator via particle swarm optimization: an application to data of collapse of Turkish commercial banks during the Asian financial crisis. J Appl Stat. 2021;48:2499-2514. [Crossref]
13. Tabachnick BG, Fidell LS, Ullman JB. Using multivariate statistics. Boston, MA: Pearson; 2007;481-498. [Crossref]
14. Seber GAF, Wild CJ. Nonlinear regression, 1989. [Crossref]
15. Başarır G. Çok Değişkenli Verilerde ayrımsama sorunu ve lojistik regresyon Analizi. Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Yayınlanmamış Doktora Tezi, 1990. [Crossref]
16. Verhulst PF. Notice on the law that the population follows in its growth. Corresp Math Phys. 1938;10:113-126. [Crossref]
17. Gürcan M, Kaya MO, Halisdemir N. Lojistik İncelemede Ayrımsama Performansının Değerlendirilmesi. Avrupa Bilim ve Teknoloji Dergisi. 2019;1008-1013. [Crossref]
18. Richards FJ. A flexible growth function for empirical use. Journal of experimental Botany. 1959;10:290-301. [Crossref]
19. Öztürk K, Şahin ME. Yapay sinir ağları ve yapay zekâ'ya genel bir bakış. Takvim-i Vekay. 2018;6:25-36. [Crossref]
20. Arf C. Makineler Düşünebilir mi ve Nasıl Düşünebilir? Atatürk Üniversitesi 1958-1959 Öğretim Yılı Halk Konferansları. 1959;91-103. [Crossref]
21. Demirhan A, Kılıç YA, İnan G. Tıpta yapay zeka uygulamaları. Yoğun Bakım Dergisi. 2010;9:31-41. [Crossref]
22. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. Applied Regression Analysis and Other Multivariable Methods, 4th ed. Thomson Learning. Inc., Belmont: California; 2007. [Crossref]
23. Kibria BG, Shukur G. On Liu estimators for the logit regression model. Economic Modelling. 2012;29:1483-1488. [Crossref]
24. Kang K, Gao F, Feng J. A new multi-layer classification method based on logistic regression. In 2018 13th International Conference on Computer Science & Education (ICCSE) (pp. 1-4). IEEE. 2018. [Crossref]
25. Eberhardt LL, Breiwick JM. (2012). Models for population growth curves. International Scholarly Research Notices, 2012. [Crossref]